Detecting repeated cancer evolution from multi-region tumor sequencing data

Detecting repeated cancer evolution from multi-region tumor sequencing data

Giulio Caravagna [ORCID 0000-0003-4240-3265] (1,2,*), Ylenia Giarratano (3,2), Daniele Ramazzotti (4), Ian Tomlinson (5), Trevor A Graham (6), Guido Sanguinetti [ORCID 0000-0002-6663-8336] (2,*) and Andrea Sottoriva [ORCID 0000-0001-6709-9533] (1,*)

1- Evolutionary Genomics & Modelling Lab, Centre for Evolution and Cancer, The Institute of Cancer Research, London SM2 5NG, UK.
2- School of Informatics, University of Edinburgh, Edinburgh EH8 9AB, UK.
3- Centre for Medical Informatics, Usher Institute, University of Edinburgh, EH16 4UX, UK.
4- Department of Pathology, Stanford University, California CA 94394, US.
5- Institute of Cancer and Genomic Sciences, University of Birmingham, B15 2TT, UK.
6- Centre for Tumour Biology, Barts Cancer Institute, Queen Mary University of London, London EC1M 6BQ, UK.

* Correspondence to: giulio.caravagna@icr.ac.uk, gsanguin@inf.ed.ac.uk and andrea.sottoriva@icr.ac.uk

Abstract

Recurrent successions of genomic changes, both within and between patients, reflect repeated evolutionary processes that are valuable for anticipating cancer progression. Multi-region sequencing allows the temporal order of some genomic changes to be inferred within a tumour, but the robust identification of repeated evolution across patients remains an unmet challenge. Here we present a machine learning method based on Transfer Learning that overcomes the stochastic effects of cancer evolution and noise in the data, and identifies hidden evolutionary patterns in cancer cohorts. When applied to multi-region sequencing datasets from lung, breast, renal and colorectal cancer (768 samples from 178 patients), our method detected repeated evolutionary trajectories in subgroups of patients, which reproduced in single-sample cohorts (n=2,935). Our method provides ways to classify patients based on how their tumour evolved, with implications for anticipating cancer evolution.

Introduction

The biggest challenge in oncology is the fact the tumours change over time, progressing from benign to malignant, becoming metastatic, and developing treatment resistance[1,2]. This occurs through a process of clonal evolution involving cancer cells and their microenvironment[3], and results in intra-tumour heterogeneity (ITH). ITH contributes to the lethal outcome of cancer by providing the substrate of phenotypic variation upon which adaptation can occur[4]. A fundamental question is therefore: can we predict a cancer's next evolutionary "step"? The question of predictability, first posed by Stephen Jay Gould for species evolution[5], is also central in oncology.

Clonal evolution results from the interplay of random mutations, genetic drift, and non-random selection[6], leading to complex patterns in the data and limiting predictability due to stochastic forces[7]. However, histopathological staging and molecular markers indicate that, at least in part, tumour evolution is predictable. Moreover, despite its stochastic nature, micro-environmental, epistatic, and lineage constraints may allow for a limited set of evolutionary steps after tumor sampling to be predicted[2]. Indeed, previous approaches based on single-sample cross-sectional data have revealed recurrent sequences of genomic events in cancer cohorts[8-11].

Recent studies have used multi-region sequencing, which allows the partial order of somatic aberrations in a tumour to be determined using phylogenetic analysis[2]. However, truncal (clonal) alterations cannot be ordered in most cases and phylogenetic trees from different patients often appear very distinct[12-18]. High levels of technical noise and biological variability currently prohibit the robust inference of repeated evolutionary trajectories across patients, despite the important implications for stratifying patients and predicting cancer progression.

Here we exploit the fact that tumours in different patients can represent multiple instances of the same evolutionary process. We devised REVOLVER (Repeated EVOLution in cancER), a method that jointly analyses multi-region sequencing data from patient cohorts by using a machine learning approach called Transfer Learning (TL)[19]. REVOLVER infers multiple patient evolutionary models jointly, with the aim of increasing their structural correlation. Our method exploits multiple independent noisy observations (i.e. single patients), and "transfers" information between patients to de-noise data and highlight hidden evolutionary patterns (Figure 1). The individual models still explain the data in each patient, while at the same time highlighting subgroups of tumours that evolved similarly.

Results

Approach and method description

Multi-region sequencing allows ITH to be assessed in individual patients, with particular focus on recurrent driver alterations. To detect repeated evolutionary trajectories across patients (Fig. 1A), the classical approach is to reconstruct the phylogenetic tree of each tumour (Fig. 1B). However, standard tools determine one tree per patient at a time, meaning that each patient model is independent and models are uncorrelated. The stochasticity and complexity of the evolutionary process, inter-patient variability and technical noise render the statistical signal of repeated trajectories very weak (Figure 1C).

A further complication is that multi-region bulk samples are cell population mixtures that require subclonal decomposition[20]. For each sample, measured allelic abundances must be transformed into the cancer cell fraction (CCF), the proportion of cancer cells carrying the mutation. However, strong tumour sampling bias confounds CCF estimates, making it difficult to infer the correct phylogenetic tree via the commonly used pigeonhole principle[21]. (The principle, which can distinguish linear and branched evolution, holds that if the CCF of two subpopulations sums to more than 1, then one subpopulation must be nested in the other; Supplementary Fig. 1.) Moreover, CCF estimation requires that sequencing data is corrected for purity, ploidy, absolute copy number, and mutation multiplicity (number of genomic copies carrying a mutation) for each variant used for phylogenetic reconstruction. This correction process propagates a significant amount of noise into the final CCF estimates and, consequently, in the associated phylogenetic trees.

REVOLVER implements a Maximum Likelihood (ML) method to *jointly* fit n models from n datasets (D_1, ..., D_n) for which either CCF or presence/absence annotations are available (Fig. 1D, Online Methods, Supplementary Note 1, Supplementary Software and https://github.com/caravagn/revolver). The method will process any alteration that can be annotated in these formats (e.g. mutation, copy number alteration, etc.). Each model is a tree that represents a partial ordering of the annotated alterations. To perform the fit, REVOLVER analyses a set of trees per patient (*solutions*) via a two-step Transfer Learning strategy that outputs n correlated evolutionary trees (T_1, ..., T_n) (Supplementary Figs. 2 and 3). Possible solutions can be pre-computed with external phylogenetic tools[22-27] and passed to REVOLVER, or can be directly computed within REVOLVER, for both CCF and binary data. The method requires a *score* per tree, which can be the model's likelihood against data, e.g., $p(D|T)$ for tree T, or any other suitable scalar that we seek to maximize.

REVOLVER uses fits to measure the heterogeneity of the trajectories, and to calculate an *evolutionary distance* to compare patients and identify tumours shaped by similar trajectories (*stratification*, Fig. 1E). Overall confidence in the predictions can be assessed with a jackknife approach[28] (Supplementary Note 2).

Finally, the genomic features of the trajectories identified using a multi-region training dataset (training set) can be exploited to classify larger, single-sample cohorts (test sets). We note that annotations of genomic features (e.g. drivers) are left to the user in order to make REVOLVER applicable to different cohorts.

Synthetic and biological validation

We validated REVOLVER against synthetic data representing 1,620 cohorts, >86,000 patients and 200,000 samples (Online Methods and Supplementary Note 3). In every test, we generated a number of random phylogenetic trees and simulated consistent CCF values from multi-region bulk profiling. True models were associated with repeated evolutionary trajectories, which we sought to retrieve with REVOLVER and standard uncorrelated phylogenetic inference. To introduce realistic allele sampling bias[21], we simulated a fraction of

cases with equally likely solutions (ambiguous CCFs, Supplementary Fig. 1). We also added Gaussian noise to model uncertainty in CCF estimates (Fig. 2A). Standard approaches use CCF data from a single patient to rank a set of possible phylogenetic trees; however, due to the uncertainty described above, the true solution does not always rank at the top (Fig. 2B), confounding the detection of repeated trajectories. REVOLVER de-noises the data and resolves ambiguity by transferring information across trees. In the presence of sampling bias, with and without technical noise, we found that REVOLVER is better at identifying the true evolutionary model, even when a large proportion of tumours have ambiguous solutions (Fig. 2C; Supplementary Fig. 4).

We next sought to validate REVOLVER against known evolutionary trajectories describing the well-studied adenoma-to-carcinoma transition in colorectal cancer[29], which proceeds via a step-wise accumulation of genomic aberrations. A significant proportion of colorectal cancers develop from adenomas, as evidenced by the success of bowel cancer screening and polypectomy procedures worldwide[30,31]. We leveraged a recent multi-region sequencing colorectal cancer dataset involving mutations in 9 adenomas and 10 carcinomas[32] (95 total samples, median 5 per patient; Supplementary Table 1, Supplementary Fig. 5). The dataset recapitulates the transition, which involves known driver genes such as APC, KRAS, TP53 and PIK3CA, and the stage of disease, adenoma or carcinoma, was hidden to REVOLVER. The method identified multiple transitions between pairs of events that characterise key evolutionary trajectories (Fig. 2D). For instance, REVOLVER leveraged information from adenomas to detect trajectories that were hidden in carcinomas (truncal mutations). The complete APC→KRAS→PIK3CA trajectory was never explicitly observed in a single patient but became detectable when patients were jointly analysed with TL. These recovered trajectories demonstrate the ability of REVOLVER to identify repeated evolution from multi-region datasets, even in cases where noise and partial observations obscure the true trajectory in most patients.

Recurrent trajectories in non-small cell lung cancer

We applied REVOLVER to the TRACERx dataset, the largest multi-region profiling effort to date, currently comprising $n = 100$ non-small cell lung cancers[18] (Supplementary Table 2, Supplementary Note 4). In this cohort, each tumour underwent whole-exome sequencing (500x depth) of multiple spatially separated regions, and a set of putative driver mutations and focal copy number alterations were annotated (302 total samples, median 3 per patient; 65421 total alterations, 450 drivers). We analysed the CCF values for all available patients ($n = 99$) and used the alterations annotated in the original study. We considered recurrent drivers those appearing in at least 2 patients, and performed a gene level analysis (i.e. we do not consider where the mutation occurs within a gene) to maximise the number of recurrent alterations. Although hotspot-level analysis could be performed, larger cohorts are required to achieve a suitable level of recurrence and transfer information across patients.

REVOLVER generated $n = 99$ correlated models and identified several repeated evolutionary transitions that grouped into 10 clusters, C1-C10 (Fig. 3A; Supplementary Figs. 6, 7). A jackknife approach[28] (Supplementary Note 4) confirmed cluster robustness, with 80% median cluster stability and strongest signal for C2, C3, C4, C6 and C8 (Supplementary Fig. 8). Clusters C4 and C6 have slightly weaker separation across resamples, and lower support is observed for small clusters like C10, or for C1 which has no clear signature. Importantly however, individual evolutionary trajectories (e.g. CDKNA→TP53) were highly robust (Supplementary Note 4). Cluster C5 describes the trajectory CDKNA→TP53→TERT (overall support >90%), suggesting progressive cell-cycle deregulation, anti-senescence, genomic instability and bypass of cell death (Fig. 3B). Two other clusters, C4 and C6, are associated with early EGFR alterations, with C4 also acquiring late TP53 loss. It is important to note that clustering the occurrences of driver alterations alone does not identify clear subgroups, even if one accounts for clonality status (Supplementary Fig. 9).

Furthermore, a comparative analysis against approaches based on single-sample cross-sectional cohorts[11], akin to refs[8-10,33,34], demonstrates the additional power of REVOLVER predictions, which combine multi-region data, phylogenetic theory and TL (Supplementary Fig. 10). By transferring information across patients, REVOLVER can also retrieve the temporal ordering of events within the same node of a tree, that could not otherwise be ordered. This feature is called expansion, and it is illustrated for patient CRUK0016 (cluster C5) where we could identify the ordering in the trunk of the tree (Fig. 3C). We also note that the phylogenetic tree fit for CRUK0016 ranked 5th out 56 possible alternatives with a standard approach, and thus would not have been inferred without TL.

Finally, repeated evolutionary trajectories extracted from multi-region sequencing data with REVOLVER can be used to derive a decision tree that classifies large single-sample cohorts. In this case, stratification of $n = 883$ single-sample tumours[35-37] demonstrate that many of the REVOLVER subgroups show significant differences in disease-free survival (Supplementary Fig. 11). Notably, previous large-scale single sample studies did not find clinically relevant subgroups using standard approaches[38].

Recurrent trajectories in breast cancer

We applied REVOLVER to a cohort of $n = 50$ primary breast cancers where multi-region whole-genome and targeted deep sequencing was available[15] (292 total samples, median 6 per patient; 403 total alterations, 296 drivers; Supplementary Table 3, Supplementary Note 4). In each sample, a panel of mutations and CNAs (cytoband-level and whole-arm) in breast cancer putative driver genes were annotated[15]. For this study, we processed all annotated mutations and CNAs as presence/absence in a sample, and considered recurrent those in at least 2 patients. REVOLVER identified several repeated evolutionary transitions (Fig. 4A) that characterised 6 evolutionary groups (Supplementary Figs. 12, 13). Again, the results were robust, but with slightly lower scores than those observed in the lung cohort possibly due to the lower resolution of binary data compared to CCF, which renders it more difficult to retrieve temporal orderings. However, the inferred trajectories were well supported by the data (Supplementary Fig. 14). For example, subgroup C2 described the repeated evolutionary trajectory TP53→PIK3CA→-8p→+8q (Fig. 4B,C), identified with >90% support (Supplementary Note 4). Again, standard clustering based on the patterns of occurrences of driver alterations does not identify similar groups (Supplementary Fig. 15).

We used repeated trajectories to create a decision tree (Fig. 5A) and stratify $n = 1{,}752$ single-sample breast cancer cases from the METABRIC[39,40] ($n = 1{,}318$) and BRCA TCGA[41] ($n = 434$) studies. We found that our evolutionary subgroups replicated in these cohorts (Fig. 5B), and survival analysis highlighted significant differences between clusters (Fig. 5C). Our evolutionary subgroups are enriched for specific breast cancer subtypes from the IntClust (based on both transcriptomic and copy number alterations) and PAM50 (transcriptomics alone) classifications (Fig. 5D, 5E). Interestingly, REVOLVER group C3, which shows significantly poorer survival and is characterised by the evolutionary trajectory TP53→+8, was enriched for IntClust 10 and basal subtypes. This analysis demonstrates how evolutionary groups identified with REVOLVER can be combined with cancer subtypes to inform how these tumours evolved.

Recurrent trajectories in renal cancer

We used REVOLVER to analyse somatic mutations in a cohort of $n = 10$ clear cell renal cell carcinomas (79 samples, median 8 per patient; 843 alterations, 75 drivers)[12]. We could identify repeated evolution involving mutations in PBRM1 and BAP1, well-known predictors of the evolution of this malignancy[12], further validating the approach. The identified trajectories reproduced in single-sample cohorts and have prognostic significance, in line with previous literature[42] (Supplementary Table 4, Supplementary Note 4).

Discussion

Detecting repeated evolution in cancer is critical for the implementation of evolutionary approaches to disease management. Stratifying patients based on their recurrent evolutionary patterns helps to predict future steps of malignant progression, thus potentially informing optimal and personalised clinical decisions.

Although the application of machine learning to biomedical datasets is becoming popular[43], the use of these methods as 'black boxes' to mine cancer genomic data is unlikely to be successful unless combined with clinical and biological knowledge. In particular, analysing results in light of the cancer evolution paradigm is essential.

Our Transfer Learning approach combines high-quality multi-region sequencing data of driver alterations and phylogenetic theory to detect the hidden signal of repeated evolution within multiple tumour types. Approaches that attempt to compare uncorrelated evolutionary models or cluster alterations fail to identify repeated evolution between patients. Our approach also helps to reconcile multi-region sequencing data with large single-sample cohorts by combining different data types and extracting more information on the evolutionary process from both strategies concurrently.

REVOLVER can be used with both binary and CCF values and can be employed in conjunction with any method providing multiple scored phylogenetic trees per patient. Importantly, our method is adaptable to a wide range of input data, making it readily usable for higher resolution datasets as they become available. Moreover, stratification power could further increase with larger datasets, and REVOLVER can be applied to single-cell sequencing data. The repeated evolutionary trajectories we identified were associated with subsets of patients with distinct prognosis, demonstrating the likely clinical value of stratifying patients based on how their tumours evolved.

Acknowledgements

This work is supported by Wellcome Trust funding jointly awarded to A.S. and T.A.G. (202778/B/16/Z and 202778/Z/16/Z respectively), as well as Wellcome Trust funding awarded to the Centre for Evolution and Cancer (105104/Z/14/Z). A.S. is supported by Cancer Research UK (A22909) and by the Chris Rokos Fellowship in Evolution and Cancer. T.A.G. is supported by Cancer Research UK (A19771). G.S. is supported by ERC (MLCS 306999).

Author's contributions

GC, GS and AS designed the approach and interpreted the results. GC defined the method. GC and YG implemented it. GC, YG and DR analysed the data. IT contributed data. GS and AS supervised the study with input from TAG. All authors drafted and approved the manuscript.

Competing interests

The authors declare no competing interests.

Figures

Figure 1. Identifying repeated evolution in cancer multi-region sequencing data using Transfer Learning. **(A)** Multi-region sampling (red circles) is used to characterise genomic intra-tumour heterogeneity (ITH). Some evolutionary trajectories are shared by patient subgroups with common somatic drivers (red or purple group) but remain hidden because of apparent variability in genomic patterns between patients. **(B)** The standard approach (top) is to infer one evolutionary model (ie, phylogenetic tree) per patient at a time, and then compare the n trees. Because models are inferred independently, statistical signal for repeated trajectories is weak and few are identified (only part of purple trajectory, bottom). **(C)** REVOLVER uses *Transfer Learning* (top) to infer n models jointly and increase their structural correlation; n trees explain the data in each patient while highlighting repeated evolutionary trajectories in the cohort (bottom).

Figure 2. Synthetic test of the method and biological validation. **(A)** Testing with synthetic data simulating 20 cohorts of $n = 50$ patients with 1-3 bulk regions each (extended tests in Supplementary Fig. 4) and modelling sampling bias in $p = 10, 30,$ or 50% of patients, as well as Gaussian technical noise ($\sigma = 0.05$). Compared to uncorrelated phylogenetic inference, REVOLVER retrieved more true trees (true positives), even for patients with ambiguous CCF data due to tumor sampling bias and noise. Boxplots show mean and inter quartile range (IQR), upper whisker is 3rd quartile $+1.5 * IQR$ and lower whisker is 1st quartile $- 1.5 * IQR$. **(B)** Biological validation using a multi-region sequencing dataset of $n = 19$ colorectal cancer patients (9 adenomas, and 10 carcinomas) covering the adenoma-to-carcinoma transition[32]. Alterations in key colorectal driver genes (rows) for every patient (columns) are shaded by the proportion of samples bearing the alteration; driver alterations are present/absent in a sample and truncal alterations are denoted by orange squares (lower heatmap). REVOLVER detected repeated trajectories (upper heatmap; e.g. APC→KRAS) which can be used to stratify patients (complete data in Supplementary Fig. 5). Distance between patients is determined from the trajectories, which contribute proportionally to their frequency in the cohort; the dendrogram is then computed by hierarchical clustering (Ward's method). REVOLVER trees (bottom) show that by transferring information across patients,

repeated evolution in early-stage tumours (adenomas) become informative of evolutionary trajectories in late-stage tumours (carcinomas), in which many alterations appear clonal and cannot otherwise be ordered.

Figure 3. Repeated evolutionary trajectories in lung cancer. (A) REVOLVER analysis of CCF data from $n = 99$ non-small cell lung cancers from the TRACERx study[18] (columns are patients). Top heatmap shows the most recurrent evolutionary trajectories (complete data in Supplementary Fig. 6). Bottom heatmap shows average CCF values (provided in[18]) for the most recurrent putative driver genes. Alterations are ordered by frequency in the cohort, truncal alterations are denoted by orange squares. REVOLVER stratified this cohort by repeated evolution into 10 evolutionary subgroups (Supplementary Fig. 7). Subgroup stability was estimated via jackknife ($N = 1{,}000$ resamples, leave out $p = 10\%$; Supplementary Fig. 8) and annotated in the dendrogram (median per cluster). These groups can be used to derive a decision-tree classifier that stratifies $n = 589$ tumours in orthogonal single-sample cohorts (Supplementary Fig. 11). **(B)** Repeated trajectories in cluster C5. Arrows indicate transitions. Number of times a transition is observed among the 9 patients, number of times an alteration is clonal or subclonal in the cohort, and probability of detecting the edge across resamples are indicated (Supplementary Note 4). **(C)** Phylogenetic model for patient CRUK0016 (cluster C5) has 13 clones (CCFs clusters), 5 with drivers annotated (in colour). The REVOLVER tree ranked 5th of 56 possibilities. Via Transfer Learning, REVOLVER can also estimate the intra-clone orderings, for example the trajectory CDKNA→TP53→TERT can be expanded.

Figure 4. Repeated evolutionary trajectories in breast cancer. (A) REVOLVER analysis of data from $n = 50$ breast cancers from Yates et al. 2015[15] (columns are patients). Top heatmap shows the most common repeated evolutionary trajectories identified by our method (complete data in Supplementary Fig. 12). Bottom heatmap shows average proportion of samples bearing the alteration (provided in[15]) for the most recurrent putative driver genes (data were presence/absence). Alterations are ordered by frequency in the cohort, truncal alterations are denoted by orange squares. REVOLVER stratified this cohort into 6 evolutionary subgroups (Supplementary Fig. 13). Subgroup stability was estimated via jackknife ($N = 1{,}000$ resamples, leave out $p = 10\%$; Supplementary Fig. 14), and annotated in the dendrogram (median per cluster). **(B)** Repeated trajectories in cluster C2. Arrows indicate transitions. Number of times a transition is observed among the 11 patients, number of times an alteration is clonal or subclonal in the cohort, and probability of detecting the edge across resamples are indicated. This group highlights the evolutionary trajectory TP53→PIK3CA→-8p→+8q. **(C)** The clone tree for patient PD14753 (cluster C2) had 11 nodes, 7 of which contain drivers (in colour). With a standard approach, this tree would have scored 2/200 alternative trees. By transferring information from other patients in the cohort (dashed lines), REVOLVER can expand evolutionary transitions within the same node. TP53 was identified as a tumour-initiating alteration (early clonal), followed by loss of 16q/17p (late clonal). Uncertainty on -16q and -17p ordering remains because of equally likely observations in the cohort. Transfer Learning also works at subclonal level, identifying the trajectory FANCD2→BRCA2. The order of MLL3 and KDR remained uncertain.

Figure 5. Stratifying single-sample cross-sectional cohorts with repeated evolutionary trajectories. (A) Subgroups identified with REVOLVER (from the multi-region breast cancer dataset in this example) can be used to build a decision tree. **(B)** The decision tree was used to classify $n = 1{,}752$ single-samples tumours from large cross-sectional cohorts (METABRIC and TCGA BRCA2012), showing that REVOLVER subgroups reproduced in large orthogonal datasets. Most recurrent driver alterations, PAM50 and IntClust classifications are reported. **(C)** Evolutionary subgroups identified by REVOLVER were prognostic (two-tailed log-rank test, $p < 0.05$, 95% confidence interval shaded). Interestingly, poor survival group C3 was enriched for a specific subset of basal tumours characterised by trajectory TP53→+8q (see Supplementary Fig. 11 for same analysis in lung cancer). **(D, E)** Enrichment of REVOLVER clusters for IntClust classification and PAM50 classifications (one-tailed Fisher's Exact test, $p < 0.05$ adjusted with Bonferroni correction, odds ratio and confidence interval in Supplementary Table 3).

References

1. McGranahan, N. & Swanton, C. Biological and therapeutic impact of intratumor heterogeneity in cancer evolution. *Cancer Cell* **27,** 15–26 (2015).
2. McGranahan, N. & Swanton, C. Clonal Heterogeneity and Tumor Evolution: Past, Present, and the Future. *Cell* **168,** 613–628 (2017).
3. Greaves, M. & Maley, C. C. Clonal evolution in cancer. *Nature* **481,** 306–313 (2012).

4. Burrell, R. A., McGranahan, N., Bartek, J. & Swanton, C. The causes and consequences of genetic heterogeneity in cancer evolution. *Nature* **501,** 338–345 (2013).
5. Gould, S. J. *Wonderful Life: The Burgess Shale and the Nature of History*. (W. W. Norton & Company, 1990).
6. Graham, T. A. & Sottoriva, A. Measuring cancer evolution from the genome. *J. Pathol.* (2016). doi:10.1002/path.4821
7. Lipinski, K. A. *et al*. Cancer Evolution and the Limits of Predictability in Precision Cancer Medicine. *Trends in Cancer* **2,** 49–63 (2016).
8. Beerenwinkel, N. *et al*. Genetic progression and the waiting time to cancer. *PLoS Comput. Biol.* **3,** e225 (2007).
9. Pathare, S., Schäffer, A. A., Beerenwinkel, N. & Mahimkar, M. Construction of oncogenetic tree models reveals multiple pathways of oral cancer progression. *International Journal of Cancer* **124,** 2864–2871 (2009).
10. Attolini, C. S.-O. *et al*. A mathematical framework to determine the temporal sequence of somatic genetic events in cancer. *Proc. Natl. Acad. Sci. U.S.A.* **107,** 17604–17609 (2010).
11. Caravagna, G. *et al*. Algorithmic methods to infer the evolutionary trajectories in cancer progression. *PNAS* **113,** E4025–E4034 (2016).
12. Gerlinger, M. *et al*. Genomic architecture and evolution of clear cell renal cell carcinomas defined by multiregion sequencing. *Nature Genetics* **46,** 225–+ (2014).
13. de Bruin, E. C. *et al*. Spatial and temporal diversity in genomic instability processes defines lung cancer evolution. *Science* **346,** 251–256 (2014).
14. Sottoriva, A. *et al*. Intratumor heterogeneity in human glioblastoma reflects cancer evolutionary dynamics. *Proc Natl Acad Sci USA* **110,** 4009–4014 (2013).
15. Yates, L. R. *et al*. Subclonal diversification of primary breast cancer revealed by multiregion sequencing. *Nat. Med.* **21,** 751–759 (2015).
16. Kim, J. *et al*. Spatiotemporal Evolution of the Primary Glioblastoma Genome. *Cancer Cell* **28,** 318–328 (2015).
17. Kim, H. *et al*. Whole-genome and multisector exome sequencing of primary and post-treatment glioblastoma reveals patterns of tumor evolution. *Genome Res.* **25,** 316–327 (2015).
18. Jamal-Hanjani, M. *et al*. Tracking the Evolution of Non–Small-Cell Lung Cancer. *New England Journal of Medicine* NEJMoa1616288 (2017). doi:10.1056/NEJMoa1616288
19. Pan, S. J. & Yang, Q. A Survey on Transfer Learning. *IEEE transactions on knowledge and data engineering* **22,** 1345–1359 (2010).
20. Roth, A. *et al*. PyClone: statistical inference of clonal population structure in cancer. *Nat Meth* **11,** 396–398 (2014).
21. Nik-Zainal, S. *et al*. The life history of 21 breast cancers. *Cell* **149,** 994–1007 (2012).
22. Schwartz, R. & Schäffer, A. A. The evolution of tumour phylogenetics: principles and practice. *Nat. Rev. Genet.* **18,** 213–229 (2017).
23. Ke Yuan, T. S. F. M. N. B. BitPhylogeny: a probabilistic framework for reconstructing intra-tumor phylogenies. *Genome Biol.* **16,** (2015).
24. Deshwar, A. G. *et al*. PhyloWGS: Reconstructing subclonal composition and evolution from whole-genome sequencing of tumors. *Genome Biol.* **16,** 35 (2015).
25. El-Kebir, M., Satas, G., Oesper, L. & Raphael, B. J. Inferring the Mutational History of a Tumor Using Multi-state Perfect Phylogeny Mixtures. - PubMed - NCBI. *Cell Systems* **3,** 43–53 (2016).
26. Salehi, S. *et al*. ddClone: joint statistical inference of clonal populations from single cell and bulk tumour sequencing data. *Genome Biol.* **18,** 44 (2017).
27. Guindon, S. *et al*. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol* **59,** 307–321 (2010).
28. Efron, B. *The Jackknife, the Bootstrap and Other Resampling Plans*. (Society for Industrial and Applied Mathematics, 1982). doi:10.1137/1.9781611970319
29. Fearon, E. R., Fearon, E. R., Vogelstein, B. & Vogelstein, B. A genetic model for colorectal tumorigenesis. *Cell* **61,** 759–767 (1990).
30. Logan, R. F. A. *et al*. Outcomes of the Bowel Cancer Screening Programme (BCSP) in England after the first 1 million tests. *Gut* **61,** gutjnl–2011–300843–1446 (2011).
31. Zauber, A. G. *et al*. Colonoscopic Polypectomy and Long-Term Prevention of Colorectal-Cancer Deaths. *New England Journal of Medicine* **366,** 687–696 (2012).
32. Cross, W. *et al*. The evolutionary landscape of colorectal carcinogenesis. *Nat. ecol. evol.*
33. Carter, S. L. *et al*. Absolute quantification of somatic DNA alterations in human cancer. *Nature Biotechnology* **30,** 413–421 (2012).

34. Prandi, D. *et al*. Unraveling the clonal hierarchy of somatic genomic aberrations. *Genome Biol*. **15,** 439 (2014).
35. Network, T. C. G. A. R. Comprehensive molecular profiling of lung adenocarcinoma. *Nature* **511,** 543 (2014).
36. The Cancer Genome Atlas. Comprehensive genomic characterization of squamous cell lung cancers. *Nature* **489,** 519–525 (2012).
37. Imielinski, M. *et al*. Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing. *Cell* **150,** 1107–1120 (2012).
38. Alexandrov, A. *et al*. Distinct patterns of somatic genome alterations in lung adenocarcinomas and squamous cell carcinomas. *Nature Genetics* **48,** 607–616 (2016).
39. Curtis, C. *et al*. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* **486,** 346–352 (2012).
40. Pereira, B. *et al*. The somatic mutation profiles of 2,433 breast cancers refines their genomic and transcriptomic landscapes. *Nat Comms* **7,** 11479 (2016).
41. The Cancer Genome Atlas. Comprehensive molecular portraits of human breast tumours. *Nature* **490,** 61–70 (2012).
42. Kapur, P. *et al*. Effects on survival of BAP1 and PBRM1 mutations in sporadic clear-cell renal-cell carcinoma: a retrospective analysis with independent validation. *Lancet Oncol*. **14,** 159–167 (2013).
43. Esteva, A. *et al*. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* (2017). doi:10.1038/nature21056

Online Methods

The number of cancer evolution studies involving multi-region sequencing are rapidly growing (see, e.g., the case studies in[12-15,17,18,44]), and intra-tumour heterogeneity profiling allows reconstructing the spatio-temporal evolutionary history of a patient tumour[2].

REVOLVER takes as input *n* multi-region sequencing datasets from *n* patients $D_1, \dots, D_n$. Each sample from a patient contains information on what genomic alterations are present in that specific sample. Our method is agnostic to the type of alteration annotated, which could be a nucleotide substitution (SNV), a copy number alteration (CNA) or any other (epi)genomic event. For each event, two data formats can be processed:

- Cancer Cell Fractions (CCF), or the proportion of cancer cells in the sample that bear the alteration.
- If CCF values are unavailable, a simpler binary format with presence/absence of the alteration in a sample.

The method also requires to specify for every patient when sets of alterations occur together in the same clone:

- For CCF data, clones are estimated via subclonal reconstruction (i.e., CCF-based clustering);
- For binary data, alterations are assumed to be in the same clone if found in the same set of samples.

For each genomic alteration, the input should also clarify if it is a putative *driver*, and/or *truncal* (i.e., present in 100% of cancer cells, or in the case of binary format, present in all samples; see Supplementary Note 1 for details on the input format).

In REVOLVER, we call alterations that are detected in multiple patients *recurrent*. We will use a parameter to determine a minimum recurrence threshold.

Evolutionary trajectories using a standard approach

For each patient, we can construct an evolutionary model (e.g. a phylogenetic tree) that explains the data via a standard approach such as those presented in refs[12-15,17,18,44]. In what follows, we will seek to compare our method to the principles underpinning those approaches.

For a cohort of *n* patients, we would identify *n* evolutionary models $T_1, \dots, T_n$ where:

- Each T_i is a tree describing the evolutionary history of a patient's tumour. Its nodes are the groups of input alterations. In the case of CCF data this is a clone tree and each node is a clone, whereas in the case of binary data this is a mutation tree[45]. The tree encodes the (partial) temporal ordering of the alterations in the tumour.
- An *evolutionary trajectory* is defined as a path $x_1 \to x_2 \to \cdots$ that connects alterations x_i, and describes their order of accumulation: x_1 is earlier than x_2, x_3, etc, while x_2 is earlier than x_3, x_4 etc. It can be computed from the ordering of the nodes in a tree.

Ideally, in order to interpret the data from a whole cohort of patients in light of tumour evolution, one would like to identify *recurrent evolutionary trajectories* describing repeated evolution across patients. Repeated evolution in cancer describes recurrent sequences of events that fundamentally underpin tumorigenesis and progression in a given subgroup of patients. Repeated evolutionary trajectories pinpoint evolutionary "steps" of a tumour, and could underlie advantageous phenotypic changes to the cancer clone.

Therefore, one needs a method that identifies trajectories that 1) are repeated across the cohort, and (hence) 2) involve recurrent alterations (drivers). Specifically, we need a method that correlates a trajectory involving recurrent drivers *x* and *y*, present within a sequence that may include passengers p_i:

$$\dots \to x \to p_1 \to \cdots \to p_w \to \mathrm{y} \to \cdots$$

See Supplementary Figure 2.

Using a standard approach based on phylogenetic theory, such as Maximum Parsimony[46] or Maximum Likelihood[27], one would infer each phylogenetic model T_i independently for each patient. A Bayesian approach would compute independently n posteriors $p(T_i|D_i)$ for $i=1,..,n$, and use them to sample models with high likelihood.

With n independent models, we could evaluate *post hoc* structural similarities between patients. However, visual inspection of a set of phylogenetic trees is impractical with complex models or large n. Automatic approaches that use structural distances, or that measure similarities among the distributions induced by these probabilistic models, can help. Nevertheless, this approach to the detection of repeated evolutionary trajectories remains impractical because cancer multi-region cohorts exhibit a high degree of heterogeneity both between and within patients (see ref[1,2] for a review), as well as inherent noise in the data.

Evolutionary trajectories using Transfer Learning

We propose a new approach to detect repeated evolutionary trajectories from noisy multi-region sequencing data of cancer patients. We assume that the recurrent trajectories can be modelled as a tree, which is *hidden* in the data. To capture heterogeneity across patients, we consider each input tumour as a noisy realisation from such tree (a realisation being the evolutionary trajectories for a patient, and its associated dataset).

In probabilistic terms, the individual patient trees are coupled through a shared prior, so that the (marginal) posterior distribution of patient trees no longer factorises across patients. Consider a *joint posterior* over $T_1, \dots, T_n$; we expect the solutions to differ in the following statistical sense

$$p(T_1, \dots, T_n|D_1, \dots, D_n) \neq \prod_{i=1}^{n} p(T_i|D_i) \,.$$

In practice, a joint inference correlates explicitly n models of evolutionary processes: the solutions will be statistically dependent, and hence correlated across patients.

We argue that the detection of statistically significant regularities from correlated models is a better approach to exploit data of n (independent) evolutionary processes that describe the same tumour. Synthetic tests show that this method improves over standard uncorrelated methods, particularly in the presence of sampling bias and technical noise in CCF.

The REVOLVER algorithm

In REVOLVER -- *Repeated evolution in cancer* -- we adopt an *Expectation Maximisation* (EM) strategy for *Maximum Likelihood* (ML) estimation of the n trees (Supplementary Note 1). The *structural correlation* among each model is measured via a parameter $\boldsymbol{w}$, which we maximise. From $\boldsymbol{w}$, we estimate repeated evolution of the n input tumours, and induce a distance metric for cohort stratification.

First, REVOLVER processes input data and group (clone) assignments to pre-compute a set of scored trees for every patient. This is done differently depending on whether CCF or binary data is available and can be modified to accommodate custom tree learning methodologies (see below).

Then, a two-steps *Transfer Learning* (TL) strategy computes the joint ML estimates of $T_1, \dots, T_n$. Very broadly, TL is a Machine Learning paradigm to exploit knowledge gained while solving multiple related tasks. Here, the inference of the model for a patient (one task) becomes informative for the inference of other models (other tasks)[19]. The *features* shared among correlated tasks are recurrent drivers and their evolutionary trajectories (i.e., orderings). We remark that TL is sometimes used to indicate a broader class of problems; in the Machine Learning literature, our approach could be more specifically called *multi-task learning*.

Precisely, REVOLVER does the following steps (Supplementary Figures 2, 3):

- computes n correlated models $T_1, \dots, T_n$, from the ones available for each patient;

- computes the evolutionary trajectories within each group of alterations annotated in every patient and refines fit estimates accordingly. These trajectories cannot be detected unless we analyse data from multiple patients, and we "transfer" trajectories across inference tasks.

REVOLVER is a model-selection strategy. We first discuss how it computes correlated models, and then how its input models can be computed from CCF or binary data.

Correlating evolutionary trajectories across patients

A dataset D_i of a single patient is a *matrix* with alterations as columns, and samples sequenced from the *i*-th patient as rows. With input CCF, each entry of D_i is a real value in [0,1]; with binary data 1s report where the alteration is detected. We assume that D_i has no **0** columns and denote as $\{D_i | i = 1, .., n\}$ the data from the whole cohort. $V = \bigcup_{i=1}^{n} V_i$ is the whole set of alterations in the cohort; V_i the ones that occur in the *i*-th patient.

Evolutionary trajectories from groups (Supplementary Figure 2). Consider a driver *x*, and denote with k_x the number of patients where it occurs; define

$$\Gamma = \{x \in V | k_x \geq \theta\} \cup \{\star\}$$

the set of recurrent alterations that occur in at least $\theta > 1$ patients, plus a special symbol $\star$ that stands for "germline" ancestor. REVOLVER processes the whole dataset and induces correlation among drivers in Γ.

We write $x \longrightarrow y \in T$ for an edge appearing in a tree T and introduce a special definition of the transitive closure of $\longrightarrow$, usually denoted as $\longrightarrow^*$ (Supplementary Figure 2,3). In general, the transitive closure of a path $x \longrightarrow y \longrightarrow z$ is the set of edges $\longrightarrow^* = \{x \longrightarrow y, y \longrightarrow z, x \longrightarrow z\}$; $x \longrightarrow z$ follows by $\longrightarrow$'s closure. In this work, we have a special interest for evolutionary trajectories among recurrent drivers. Consider for the *i*-th patient the trajectory

$$p'_1 \longrightarrow \cdots \longrightarrow p'_z \longrightarrow x \longrightarrow p_1 \longrightarrow \cdots \longrightarrow p_w \longrightarrow y \longrightarrow \cdots \quad \text{where } p_i, p'_i \notin \Gamma \text{ and } x, y \in \Gamma$$

We write $\pi_y^i = x$ to denote the recurrent driver upstream of *y* in this patient; these trajectories are correlated in REVOLVER. We indicate them by the notation $\pi_y^i = x \longrightarrow^* y$, or when it is clear by $x \longrightarrow^* y$.

Because input alterations are grouped into clones, we need to account for groups when we create trajectories. If $g_1 \longrightarrow g_2$ are two groups in a model's path, and x_j and y_j are the driver alterations in those groups, we account for all combinations of orderings in the two groups with the trajectories

$$g_1 = \{x_1, \dots, x_w\} \longrightarrow g_2 = \{y_1, \dots, y_l\} = \begin{cases} x_1 \longrightarrow y_1 \\ \cdots \\ x_w \longrightarrow y_l \end{cases}$$

This creates a combinatorial number of trajectories according to the number of drivers annotated in each group of a patient's alterations. Clearly, the trajectory within a patient's group is a linear ordering of its alterations that, however, cannot be estimated from a single patient. This is a confounding factor that renders the inference harder. However, by leveraging cross-sectional data from multiple patients diagnosed at different evolutionary times, one can recovery such trajectories and average out the confounders.

Multinomial counts of trajectories. To measure the structural correlation among the models, we count how often they contain a path that connects *x* and *y* in Γ; the minimum among k_x and k_y is an upper bound to this count.

Definition (Multinomial consensus) *Given n* trees $T_1, \dots, T_n$, *we define the* $|\Gamma| \times |\Gamma|$ *discrete-valued* consensus matrix $\boldsymbol{w}$ *with entries*

$$\boldsymbol{w}_{x,y} = |\{T_i \mid x \longrightarrow^* y \in T_i; x, y \in \Gamma\}|$$

where $x \longrightarrow^* y$ *is a trajectory defined as explained above* (Supplementary Figure 3).

Clearly, $\boldsymbol{w}_{x,y}/k_y$ is an empirical probability for the observation of *x* upstream *y* in the *n* models. By construction, we are detecting a statistical signal among *x* and *y*, recurrent driver alterations that intertwine with passengers. The role of $\star$ is to capture which $x \in \Gamma$ is earliest in the trunk of a model (the associated trajectory is $\star \longrightarrow^* x$); so $\boldsymbol{w}_{\star,x}$ counts how many tumours are predicted to initiate via *x*. It must follow by tree construction that no alteration is upstream $\star$, and hence $\boldsymbol{w}_{x,\star} = 0$.

Model-selection via Transfer Learning. REVOLVER requires a pre-computed set of trees per patient, and their scores (that must be sortable values); the algorithm uses those sets of models and $\boldsymbol{w}$ as estimator of their structural correlation and selects each patient's most correlated tree. Procedures to create trees are implemented in the framework, according to the input data; see Supplementary Note 1, for the algorithms' pseudocode.

REVOLVER*'s score* of a model T_i is a rescaling of its pre-computed score by a factor that measures its structural deviation from the models of the other patients. The pre-computed score acts as a log-likelihood of the data under the model: $p(T_i|D_i)$.

Definition (Model's score) *Let* $\Gamma_i = V_i \cap \Gamma$ *be the recurrent drivers in patient i. A model* T_i *for this patient has* score

$$f_{T_i} = \log p(T_i|D_i) + \log p(T_i|\boldsymbol{w})^{\alpha}$$

for $\alpha \geq 1$. *The latter term is a* regularization term

$$p(T_i|\boldsymbol{w}) = \prod_{x \in \Gamma_i} \left(1 - \frac{\sum_y \boldsymbol{w}_{y,x} - \boldsymbol{w}_{\pi_x^i,x}}{k_x}\right).$$

If the pre-computed scores factorize over models' edges, we can decompose the score as

$$\log p(T_i|D_i) \propto \log \prod_{x \longrightarrow y \in T_i} p(y|x; D_i)$$

where $p(y|x; D_i)$ are the edge terms obtained by fitting the tree's parameters to D_i. This factorization is common but is not a requirement. Technically, f_{T_i} is a penalised log-likelihood; we refer to $1 - p(T_i|\boldsymbol{w})$ as the *penalty* that re-scales T_i's likelihood at polynomial rate with degree α. This overall quantity is the "*information transfer*" (Supplementary Figure 2); α is a scaling factor that "shrinks" the penalty effect; in practice we always set it to 1 but it could be easily used to induce a stronger effect of the information transfer in shaping the gradient. In Supplementary Note 2, we show power calculations for the minimum information transfer to induce an ordering's swap.

We observe the following properties of the above definition:

I. the information transfer considers only penalties by predictions that disagree with T_i. In fact, for any $\pi_x^i \longrightarrow^* x$ in T_i, term $\boldsymbol{w}_{\pi_x^i,x}$ is subtracted from the penalty;
II. we penalise independently each recurrent driver $x \in \Gamma_i$, proportionally to the consensus of its evolutionary trajectories $\sum_y \boldsymbol{w}_{y,x}$ across the cohort;
III. $\star$ does not have incoming edges; only its outgoing edges contribute to $p(T_i|\boldsymbol{w})$.

Definition (Model selection) *To select n models* $\mathbf{T}_* = [T_1, \dots, T_n]$, *we solve a problem of discrete optimisation*

$$\mathbf{T}_* = \arg\max_{\boldsymbol{T}=[T_1,\dots,T_n]} [\, f_{T_1}, \dots, f_{T_n}]$$

This problem is approached with an EM procedure. Because the trees are pre-computed for each model, a global solution for each initial EM condition is guaranteed. Given an initial estimate of the trees, $\mathbf{T}^{(0)}$ we compute $\boldsymbol{w}^{(0)}$ to select the $\mathbf{T}^{(1)}$ that maximise REVOLVER's score under $\boldsymbol{w}^{(0)}$. We then iterate by estimating $\boldsymbol{w}^{(1)}$from $\mathbf{T}^{(1)}$, etc.; we stop when we reach a fix-point $\mathbf{T}^{(i+1)} = \mathbf{T}^{(i)}$ for some *i*, which is the ML estimate of $\mathbf{T}_*$.

Precisely, the E and M steps are (Supplementary Figure 3):

- [E-step] from the current estimates of $[T_1, \dots, T_n]$, compute $\boldsymbol{w}$*;*

- [M-step] use $\boldsymbol{w}$ to compute the penalty; for every patient update the scores of its pre-computed models, and determine the highest scoring (ML estimate).

The ML estimates push to minimise the penalties in the sense that the optimisation gradient pushes $p(T_i|\boldsymbol{w})$ to 1. In fact, with penalty 1 all the models predict the same trajectories for the variables in Γ, and we reach the objective of maximising the number of models, out of k_x, that predict the same driver upstream x. To start this EM, one can sample multiple random initial conditions, and select the solution with lowest penalty; this can be done in parallel. Equivalence classes of solutions with the same score and penalty might exist; this depends on the distribution of the input pre-computed scores. The method, however, is more powerful than its un-correlated counterpart in estimating the true model, as we measured via synthetic tests.

Computing trajectories within groups (expansion)

We know that, in every model T_i, we cannot compute trajectories for the alterations $x_1, \dots, x_w$ that map to the same group g (e.g., those in the same clone). However, their trajectories might be detectable in those patients $T_j \neq T_i$ whose alterations overlap with g, if they are sampled at an earlier time. Because the hidden model is assumed to be the same tree for all patients, T_j's trajectories are representative of the ones hidden in T_i.

In a TL approach, we transfer this information to T_i and split g accordingly; we can do that once the first EM strategy has converged. We call this procedure "expansion" of a group (Supplementary Figure 3). This heuristic first subsets the entries of $x_1, \dots, x_w$ from $\boldsymbol{w}$, and then selects, for each x_i, the most frequent parent driver. This is the multinomial ML estimate in $\boldsymbol{w}$; if this does not exist because there is no evidence of any of the drivers in g to be upstream x_i, then x_i cannot be ordered and will be associated to the node upstream g. Ideally, if the input tumours were homogenous and we add observations from patients at different steps of progression, we could retrieve the unknown linear ordering (i.e., a topological sort) of $x_1, \dots, x_w$. In realistic cases, because of the uncertainty in the estimation of these trajectories and drivers' annotation, we expect the expansion to be a graph that, of course, does not represent branched evolution.

Notice that the expansion does not change T_i's original likelihood (since its data was uninformative of g's trajectories), but it still changes the tree structure, and hence $\boldsymbol{w}$ and the penalty. We expect expansion to reduce the variance of $\boldsymbol{w}$; if the cohort were truly homogenous, the penalty should decrease as well since we are selecting one particular ordering of $x_1, \dots, x_w$ from a homogenous cohort.

Building input models from CCFs

Consider a patient with c groups – in what follows called clones for consistency with CCF-based studies – from r sequencing samples, its CCF data is stored in a $c \times r$ real-valued a matrix M. Each entry is a value in $[0, 1]$, estimated from read counts, the input clone assignments of each alteration, copy number segments and tumour purity. REVOLVER's implementation provides a method to compute phylogenetic trees to use as input for the tool. The tool allows one to input also a custom set of trees and scores. See also Supplementary Note 2.

Generating trees. The method implemented exploits a modified version of ClonEvol, a tool for phylogenetic inference from CCF clusters[47]. This tool first enumerates, independently for each sample, all trees compatible with M and rooted in z, the truncal clone. Then, it tries to build a "consensus" tree model that fits all the r regions at once. To build a tree, ClonEvol uses the standard *pigeonhole* principle[21]: for a node x to branch towards $y_1, \dots, y_k$, the parent's CCF must be greater than the sum of y_i's CCF, that is

$$\mathbf{ccf}(\mathrm{x}) > \sum_{i=1}^{k} \mathbf{ccf}(y_i).$$

Clearly, certain combinations of CCF values are ambiguous, and support alternative trees. For instance, if x has CCF 1 and y and z 0.3 and 0.1, then both the linear path $x \rightarrow y \rightarrow z$, and the branched model (x towards y and z) are plausible under the pigeonhole principle. Because of noise in CCF estimation and tumour sampling bias, a consensus model might only be available if we allow for violations of such principle.

Ranking phylogenetic trees. We are not interested in a perfect consensus model, but rather we want to generate several alternative trees to input to REVOLVER. We modified ClonEvol to skip its last step and return the trees computed per region. With that, we could create a *distribution of trees* plausible under the input CCF, with a probability mass proportional to the extent to which a tree violates the pigeonhole principle under M, and the

empirical evidence of each edge (obtained from ClonEvol estimates). This ensures that, even without perfect CCF, we can still compute a model for the data, and quantify its goodness of fit, without sub-setting input.

We proceeded as follows. Consider *C*, the set of clones annotated in *M*, and merge all trees into a *weighted direct acyclic graph D* whose nodes are *C*, and the weights are the average frequency of detection of the edges in each region, as estimated in ClonEvol. For each edge $x \to y$, this is the empirical probability $\lambda_{x,y}$ of clone *x* to be a direct parent of *y* in the phylogenetic trees, according to the trees estimated by ClonEvol. Thus, *D* is a generator of the distribution of phylogenetic trees for data *M*, assuming all edges to be independent.

The support of this distribution is the set of all minimum-spanning trees rooted in the truncal clone, which is known. This can be generated exhaustively only for small number of clone $c = |C|$, i.e., for a few thousand trees. If this is not the case, we can Monte Carlo sample a desired number of distinct trees for this patient; for each node *y*, its parents are sampled from the discrete marginal distribution $\boldsymbol{\lambda}_y = \{\lambda_{x,y}\}$. This exploits a factorization of the distribution over the tree's nodes and leads to sample trees that maximize the observed frequencies of edges, as we might desire.

Definition (Phylogenetic score) *For a set of phylogenetic trees* $\mathcal{T}$*, each* $T \in \mathcal{T}$ *can be scored as*

$$\eta(T) = \prod_{x \in T} \epsilon(x) \prod_{x \to y \in T} \lambda_{x,y} \qquad \epsilon(x) = \frac{1}{r} \sum_{i=1}^{r} \mathbf{1}_{\mathbf{ccf}}(x, i)$$

.

where $\mathbf{1}_{\boldsymbol{ccf}}$ *is an indicator function that evaluates to 1 if x satisfies the pigeonhole principle in the i-th region, and 0 otherwise*.

This score has the following desirable properties:

- $\eta(T)$ and $\epsilon(x)$ span in [0,1], and allow for equivalent-scoring models;
- $\epsilon(x)$ is a goodness-of-fit measure: lower values indicate increasing violations of the principle, for *x* in *T*, under data *M*.
- terms $\lambda_{x,y}$ is a probability that measures how often ClonEvol predicts *x* upstream of *y*; when this approaches 1 we have stronger evidence that *x* is upstream of *y*.
- $\eta(T) = 1$ only when 1) there is a unique possible assignment to the parents of every clone, and 2) there are no violations of the pigeonhole principle.

This score $\eta(T)$ is a *joint likelihood*: the probability of each parent of a clone is weighted by a multinomial likelihood of error $\epsilon(x)$ estimated from the tumour data. This part of the algorithm can accommodate several customizations, and it is straightforward to use phylogenetic tools that provide alternative scoring function[22-27].

For our score or variations thereof, the following min/max interpretation holds. If we maximize $\eta(T)$ alone we select the tree with most-frequent structure (*max*), and the smallest violations (*min*). When $\eta(T)$ is combined within REVOLVER we expect a *min/max-max* shrinkage effect where, at the same time, we minimize errors in each phylogeny, and maximize both tree edges that are frequent *and* represent repeated evolution in the cohort.

Building input models from binary observations

Binary data is lower-resolution than CCFs but can still be used to create a mutation tree for a patient. To do that, REVOLVER implements a method that links Suppes' theory of probabilistic causation to cancer progression[11,48,49].

Definition (Suppes' probabilistic causation in cancer) *For any two variables x and y, edge* $x \longrightarrow y$ *can exist in Suppes'* probabilistic *model only if* $p(x) \geq p(y)$ and $p(y|x) \geq p(y|\neg x)$, where $p(\cdot)$ are empirical multinomial probabilities estimated *via ML from binary data*.

A Suppes' *partially ordered set* (poset) Π_i is the set of edges that satisfy probabilistic causation. We estimate for patient *i* its poset by data D_i, and use it as building blocks of our mutation trees. Temporal priority acts as both an *infinite sites assumption*, and a *no back-mutations* model (in phylogenetic jargon). In practice, we are assuming that alterations are persistent and, accordingly, we estimate temporal precedence via marginal frequencies. Probability rising, instead, is a measure of the degree of association between two variables, which

implies statistical dependence as it is symmetric (like correlation), see Supplementary Note 2 for further discussion.

A poset is also a weighted directed graph with constant normalized weights, if we assume all poset's parents equally likely. So, it can be used to generate all minimum spanning trees rooted in the clonal group, which is the one whose alterations appear in all samples. Mutation trees can be sampled as done for phylogenetic trees, either exhaustively or by Monte Carlo, and can be scored via standard information theory. Each such model is a well-known Chow-Liu tree, a generator of the joint distribution $p(c_1, \dots, c_w)$ if $c_1, \dots, c_w$ are the w groups for this patient – i.e., the probability of observing the presence/ absence of the corresponding alterations in a sample[50]. A Chow-Liu tree contains second-order terms $p(y|x)$ for the product approximation of the joint distribution that we factorise. It is well known that it has the minimum Kullback-Leibler divergence to the true distribution, being its closest approximation in an information-theoretic sense.

Definition (Binary tree score) *For a set of Chow-Liu trees $\mathcal{T}$, each $T \in \mathcal{T}$ can be scored as*

$$\tau(T) = \prod_{x \to y \in T} \mathbf{mutinf}(x, y)$$

.

where $\mathbf{mutinf}(x, y)$ *is the* mutual information *associated to random variables that take values x and y*

$$\mathbf{mutinf}(X, Y) = \sum_{x,y} \frac{p(x, y)}{p(x)p(y)} \quad .$$

Thus, the highest-scoring Chow-Liu tree is the optimal solution to this model-selection task. REVOLVER's input Chow-Liu trees can be ordered by decreasing mutual information; our method will fit lower-rank ones only if they have smaller penalty.

Synthetic tests

We carried out synthetic tests with CCF data to validate and assess the performance of REVOLVER under different configurations of cohort size, number of samples per patient and other covariates modelling confounding factors. Tests and results are detailed in Supplementary Note 3 and Supplementary Figure 4.

In a first batch of tests, we generated phylogenetic trees and CCF data under a combined model of *tumour sampling bias*. Statistically speaking, in some patients CCF will be hard to process (i.e., noisy): they will suggest linear and branched models of evolution with the same score. In other patients, CCF data will top-score the true evolutionary model. Results show that REVOLVER, by transferring information across patients, can retrieve the true model also for patients with noisy CCF data. Uncorrelated inference (the baseline method that we compare against), instead, suffers from sampling bias and uncertainty in tree estimation. This shows that joint ML estimation of the correlated trees can de-noise genomics data, improving on the uncorrelated counterpart.

In a second batch of tests, we investigated resistance to noise of our estimator. REVOLVER's information transfer is estimated from data, thus if CCF data are dominated by noise, the algorithm will transfer "noise" and might fit repeated errors. We investigate this phenomenon with synthetic datasets affected by different intensities of Gaussian noise (technical noise) and show that REVOLVER is robust for reasonable ranges of those parameters.

Further material and case studies

REVOLVER is a framework with other features beyond its main inferential algorithm.

In Supplementary Note 2 and 4 we present:

I. Power calculations to correlate evolutionary trajectories;
II. A scalar index of divergent evolution that measures the heterogeneity of the trajectories inferred;
III. A REVOLVER-derived evolutionary distance (grounded in ecological theory for species' diversity) to stratify the cohort into subgroups of tumours that harbour similar evolutionary trajectories.

IV. A jackknife approach to estimate the stability of clusters and trajectories.
V. Further commentary on the approach;
VI. Algorithmic settings for the analysis of real data.

Data availability

REVOLVER is available as an open source R tool at https://github.com/caravagn/revolver – a copy of the source code is available enclosed with this manuscript. The datasets used in our analyses have been downloaded from the corresponding publications, and are also available alongside the tool. The source code to replicate all our analyses is available in the form of RMarkdown vignettes available at the tool's webpage.

References (Online Methods)

44. Gerlinger, M. *et al*. Intratumor Heterogeneity and Branched Evolution Revealed by Multiregion Sequencing. *New England Journal of Medicine* **366,** 883–892 (2012).
45. Davis, A. & Navin, N. E. Computing tumor trees from single cells. *Genome Biol*. **17,** 86 (2016).
46. Swofford, D. L. PAUP*: Phylogenetic Analysis Using Parsimony (and Other Methods) 4.0 Beta. *Sinauer, Sunderland, MA* (2005). Available at: http://www.sinauer.com/paup-phylogenetic-analysis-using-parsimony-and-other-methods-4-0-beta.html.
47. Dang, H. X. *et al*. ClonEvol: clonal ordering and visualization in cancer sequencing. - PubMed - NCBI. *Annals of Oncology* **28,** 3076–3082 (2017).
48. Loohuis, L. O. *et al*. Inferring Tree Causal Models of Cancer Progression with Probability Raising. *PLoS ONE* **9,** e108358 (2014).
49. Ramazzotti, D. *et al*. CAPRI: efficient inference of cancer progression models from cross-sectional data. *Bioinformatics* **31,** 3016–3026 (2015).
50. Chow, C. & Liu, C. Approximating discrete probability distributions with dependence trees. *IEEE Trans. Inform. Theory* **14,** 462–467 (1968).